E401: Empirical Challenge

Data Visualization

Fall 2022

September 2022

Please work on this challenge with a partner. All challenges are based on real-world data and are similar to what you might encounter in your future career. The main purpose of this challenge is not on the application, but on getting you (and your class mates) more familiar with applying the techniques that we discussed in the lecture. Nevertheless, whatever you present should make economic sense and if it doesn't you should think about what could be going wrong with the data and/or your analysis. Please keep in mind that these are (potentially dirty) real-world data and I haven't checked every detail of it. Therefore, you are likely to run into a lot of problems. I strongly encourage you to come to my office hour to discuss any issues as well as your overall plan for your presentation a few days before the respective class. There is always a risk that there is not much interesting in your data set. As long as you are able to clearly document what you tried and have some conjecture/explanation for why you get the results you get, this is totally fine. It's very likely that you will be in similar situations regularly when taking a job as a data scientist. I designed this challenge to be pretty open-ended on purpose. When diving into the data you may find aspects that are totally different from what I had in mind. This is totally fine and another likely outcome in data science projects.

You are expected to give a presentation of roughly 25-30 minutes in class. Think of this presentation as one you would give to your boss or at a board meeting of a company or policy institution that hired you as a data scientist. Other students should think of themselves as board members who attend your presentation and are strongly encouraged to ask critical questions about your analysis and you should be prepared to answer them. Your presentation should contain the following elements: (1) a brief discussion of the data, i.e., where is it coming from, what are the most important variables, what is the unit of observation, what concerns do you have about the quality of the data etc., (2) the big picture business or policy question that you are trying to address with these data (other students have not necessarily read the questions in advance), (3) overview of the methodology that you used to answer the question, (4) your empirical results, (5) discussion of the results, policy implications, and potential caveats and suggestions for further steps. Lastly, this is not a presentation class, so don't invest in fancy PowerPoint slides! Having prepared a RScript in RStudio that generates all your results as we click through it is totally fine! However, I ask you to only work with code scripts. Avoid manual manipulation or loading of the data from a graphical interface at *all costs!*

Main Techniques

In this challenge I will ask you to work mostly with data visualization techniques.

Data

In this challenge you will use information on the product characteristics of various ready-to-eat breakfast cereals. The data comes from this website: https://perso.telecom-paristech.fr/eagan/class/igr204/datasets.

Some variables deserve further explanation:

1. mfr: Manufacturer of cereal A = American Home Food Products G = General Mills K = Kelloggs N = Nabisco P = Post Q = Quaker Oats R = Ralston Purina
2. type: c = cold cereal h = hot cereal
3. shelf: typical display shelf in supermarket (1, 2, or 3, counting from the floor)
4. weight: weight in ounces of one serving
5. cups: number of cups in one serving
6. rating: a quality rating of the cereals from Consumer Reports

The other variables should be self-explanatory, but if you're in doubt, please get in touch with me.

Before you run any analysis, make sure you familiarize yourself with the data and examine its quality. Briefly mention in your presentation, if some features look dubious to you.

Business question

PostNabisco, one of the largest cereal manufacturers in the US, has hired you as a consultant.[1] The RTE breakfast cereal market has been in decline for several years and PostNabisco is thinking about refreshing its product lineup by introducing new cereals. In particular, they are considering building up a stronger presence in the market for health-conscious consumers.

Before they start the product development they ask you for a thorough market analysis. However, none of the board members or division directors have ever taken an Econometrics class, so they ask you to conduct your analysis exclusively with visual tools (graphs). Unfortunately, they are not very specific in their instructions and all they provide you with is a series of loose thoughts that they brainstormed:

[1] Even though Post and Nabisco are two different brands, they are actually one company.

- How can we generally classify cereals currently on the market? What are their key characteristics? Can we identify different types/groups of cereals?
- In the past we had good experiences with entering in segments of the market that are “less crowded”. For example, if we introduced a new product that is high in fiber and other healthy ingredients and nobody else is selling such a product, we would essentially have a monopoly on that part of the market.
- Given the above reasoning is it a good idea to develop a new “healthy” cereal or should we rather target another market segment?
- Some of our shareholders also have stakes in other cereal manufacturers and we want to see into which rival firms’ market share we would likely cut most by introducing a new product. Can you tell us if our competitors specialize in certain product types?

After being frustrated with the very vague requests and the pretty small data base, you get to work:

1. Try to answer the board’s question using a series of graphs that visualize the most relevant aspects of your data. Remember that you only have 30-40 minutes including questions; therefore, you may not get to answer every one of their questions. Your presentation should be flexible enough to accommodate running out of time, but you should be able to say at least a little bit about each of their questions.
2. Clearly explain how you constructed your graphs and how you interpret them. Be specific. A lot of the managers are easily distracted and unless you are very precise they are likely to misunderstand you. Assume that your audience has a solid business background, but they do not necessarily remember their statistics courses very well.
3. What would your final recommendation for a new product introduction be? Explain clearly how you arrived at your recommendation and what some of its limitations are. What would be the most important thing you need to obtain or do in order to overcome some of the limitations?
4. Finally, think about what else PostNabisco might be able to learn from the data. What future steps would you suggest to take?